Automatic Summarisation: 25 Years On

CONSTANTIN ORĂSAN
Research Institute in Information and Language Processing,
University of Wolverhampton, UK
e-mail: {C.Orasan}@wlv.ac.uk

(*Received 11 May 2013; revised 07 August 2013*)

Abstract

Automatic text summarisation is a topic that has been receiving attention from the research community from the early days of computational linguistics, but it really took off around 25 years ago. This article presents the main developments from the last 25 years. It starts by defining what a summary is and how its definition changed over time as a result of the interest in processing new types of documents. The article continues with a brief history of the field and highlights the main challenges posed by the evaluation of summaries. The article finishes with some thoughts about the future of the field.

1 Introduction

Automatic text summarisation is one of those research topics that has been around since the early days of computational linguistics and is still receiving a lot of interest from the research community, as it is far from being considered a solved problem.[1] As the name suggests, the purpose of automatic text summarisation is to develop automatic methods that take one or several texts and produce a summary from them. Over years both the type of input texts, which will be referred to as the *source*, and how the expected summary should look, referred to as the *output*, have influenced the research carried out in the field.

The purpose of this article is to provide a quick overview of the main developments in the field of automatic summarisation with emphasis on the those that have taken place in the last 25 years, since the first issue of the Journal of Natural Engineering (JNLE) was published. Given that this is the anniversary issue, I refer to relevant publications from JNLE to illustrate the evolution of the field as much as possible. Due to space restrictions, the purpose of this article is not to provide a comprehensive survey of the field. For researchers keen to find out more there are several books and survey articles which give a very detailed overview. Mani and Maybury (1999) is a collection of important articles on automatic

[1] In this paper, I use the terms *Computational Linguistics* and *Natural Language Processing* interchangeably. I also use the terms *automatic summarisation* and *text summarisation* to mean *Automatic text summarisation.*

summarisation which were previously published in various places, but at that time were not easily accessible to researchers. The most authoritative book written in the field is (Mani, 2001a). Both publications are now somewhat out of date given that their focus is on research carried out before 2000, but they give a very good overview of the field when JNLE started. Survey articles such as (Saggion, 2008; Lloret and Palomar, 2011; Nenkova and McKeown, 2012; Yao, Wan, and Xiao, 2017; Lloret, Plaza, and Aker, 2018; Gupta and Gupta, 2019) provide detailed and up-to-date information on the field.

One of the important questions that needs to be answered before attempting to implement any method for producing summaries is *"What is a summary?"*. This paper starts with Section 2 which shows how the definition of a summary has evolved over time, influencing the research carried out in the field. Section 3 presents a brief history of the main stages of development in automatic text summarisation. The availability of a well-established evaluation methodology, as well as the necessary resources to carry out the evaluation, is instrumental to the progress in any research field. For this reason, Section 4 discusses the main evaluation approaches used in automatic summarisation. The paper concludes with some thoughts and wishes for the future of the field.

2 What is a summary?

The original purpose of automatic text summarisation was to enable computers to produce summaries that are on par with those written by human professional summarisers. However, researchers working in the field quickly realised that summaries that follow standard definitions such as "an abbreviated, accurate representation of a document which should be published with it and which is also useful in secondary publications and services" (ANSI, 1977), or "an abstract summarises the essential contents of a particular knowledge record, and it is a true surrogate of the document" (Cleveland, 1983)[2], cannot be produced using computers. As a result, work carried out in the late 1990s proposed less ambitious definitions such as "a concise representation of a document's content to enable the reader to determine its relevance to a specific information" (Johnson, 1995).

Fields such as machine translation (MT) have quite a well defined goal: to develop programs that can translate from one language to another. For this reason, the focus of research in machine translation was to have increasingly better translations. In contrast, the goal of automatic summarisation kept moving as a result of changes in the types of documents to be processed and types of summaries that had to be produced.[3] This was due to the interest in processing new types of documents

[2] In the field of professional summarisation the terms *summary* and *abstract* are usually used as synonyms. In automatic summarisation, abstracts are one type of summary.

[3] I am aware that this is an oversimplification of the research carried out in machine translation, which may upset some researchers in that field. The point here is that incremental research was possible in MT. In contrast, as a result of changes in the focus of automatic summarisation, at times it was necessary to start from scratch as if working in a brand new field.

and an increasing confidence in the capabilities of NLP methods. The purpose of this section is to discuss different definitions for summaries. Section 2.1 presents a taxonomy to classify different types of summaries proposed by Spärck-Jones (1999). The main types of summaries produced by researchers in the field are presented in Section 2.2.

2.1 Characteristics of summaries

When producing summaries researchers have to consider the documents that have to be summarised and how the output should look. These are important for deciding which are the most appropriate methods for producing the summaries and how their output should be evaluated. This fact was captured by the taxonomy developed by Spärck-Jones (1999) who proposed the use of *context factors* to define characteristics of summaries. According to her classification, the factors which influence summaries can be divided into three categories: *input factors*, *purpose factors* and *output factors*. The input factors characterise the input document(s) in terms of *structure*, *genre*, *language*, *format* and *unit*. The purpose factors indicate the relationship between source and output. These factors are the most important ones among the context factors, because they have the greatest influence on choosing a summarisation method. The purpose factors identified by Spärck-Jones (1999) are *situation*, *audience* and *use* and describe who will use the summary, and in which way it will be used. The output factors define the output of the system (i.e. in this case the summaries), and are largely driven by the purpose factors. Although intended for automatic summarisation, these context factors can be easily adapted to other fields in computational linguistics.

2.2 Types of summaries

Over time, the definition of a summary has changed depending on the different values for the context factors. In some cases there are different values for the input factors (e.g. the difference between single- and multi-document summarisation), but in other cases it depends on the audience (e.g. generic- vs user-focused summaries). Each type of summary requires a different approach to producing summaries. In some cases, the differences between the approaches are not too great (e.g. it is possible to adapt methods which produce generic summaries for user-focused summarisation), but in some instances it is not uncommon to have completely different approaches (e.g. extracts vs abstracts).

2.2.1 Single-document vs multi-document summarisation

The definitions listed above, as well as most of the early work in text summarisation, focused on *single-document summarisation*, which involves taking information for the summary from only one document. Historically, this was driven by the need to have ways to produce summaries from scientific articles in an attempt to improve access to them (Kupiec, Pedersen, and Chen, 1995; Teufel and Moens,

2002). Single-document summarisation methods for use with newswire texts were also developed, but in many cases the lead summary (i.e. the first few sentences of the news article) is a very difficult baseline to beat. Most of the methods developed during the revival of the field (Section 3.3.1) focused on single-documents summarisation.

The increasing amount of information available in electronic format and the developments in the field of computational linguistics led to a diversification of the types of texts processed and the types of summaries produced, in turn leading to alternative definitions of what a summary is and how it should be produced. For example, a significant part of the work carried out since 2000 has been on multi-document summarisation and for this reason Hovy (2003) states that "a summary is a text produced from one or more texts, that contains a significant portion of the information in the original text(s)".

The move from single-document summarisation to multi-document summarisation was largely promoted by the need to deal with the increasing amount of information available online. Initially, the methods were applied to newswire texts, but more recent work focused on summarisation of user-generated content such as discussions on forums (Tigelaar, Op Den Akker, and Hiemstra, 2010; Verberne, Krahmer, Wubben, and van den Bosch, 2019), customer reviews with the output focused on product features to assist customers' decision making (Feiguina and Lapalme, 2007).

Whilst the main challenges in single-document summarisation were determining the most important information and production of a coherent summary, multi-document summarisation brought a brand new set of challenges such as a much higher compression rate, and the need to anchor sentences on the timeline in order to present information in chronological order. Redundancy of information is another challenge that needs to be addressed in multi-document summarisation, which becomes even more problematic for user-generated content such as discussions from forums because they contain a significant amount of repeated text. Given that multi-document summarisation methods usually have to deal with a large amount of input data, the best way for visualising the output should also be considered (Ando, Boguraev, Byrd, and Neff, 2005).

2.2.2 Extracts vs. abstracts

Summaries produced automatically can be *extracts* or *abstracts*. Extracts contain units from the document (i.e. paragraphs, sentences or clauses) which are used without any modification, whereas abstracts include units which are not present in the document and which are obtained using methods such as aggregation, deletion, or generation (Hovy and Lin, 1999).

The vast majority of work carried out in text summarisation has focused on extractive summarisation. This is not surprising given that this process requires only to identify the most salient sentences for a summary and present them as they are (in many cases in the order they appear in the source). This approach has some limitations because in many cases not all information present in a sentence is

relevant for a summary. In contrast, abstractive summarisation produces text which include units that are not present in the source document(s). In order to avoid complicated processing like text understanding and generation, as is the case with the methods presented in Section 3.2, researchers produced abstracts by identifying sentences that contain important information and then fuse them into new sentences (Barzilay and McKeown, 2005) or compress them Knight and Marcu (2002). Recent research in deep learning for automatic summarisation has proposed new methods to produce abstracts (see Section 3.3.3). The headlines generated by summarisation methods are another example of automatic abstracts. These were common during the Document Understanding Conference (DUC) evaluations (see Section 4.1) and re-emerged as a success of the neural approaches discussed in Section 3.3.3.

2.2.3 Generic vs user-focused summaries

Most of the single-document summaries are also *generic summaries* which means that they try to cover all the topics mentioned in the source documents. Given the variety of topics covered in a collection of documents that have to be summarised, multi-document summaries are normally *user-focused summaries* (also referred to as *query-based summaries*). This means that they focus on one topic or set of topics that are relevant for the reader of the summary. The most common way to define this topic is as a query that is employed both to retrieve documents using an information retrieval engine and to ensure the focus of the summary produced from these documents (Goldstein, Kantrowitz, Mittal, and Carbonell, 1999). The user's interest can also be defined as a question that is sent to an advanced question-answering system which produces its answers by fusing information from several sources. For example, Verberne, Boves, Oostdijk, and Coppen (2010) propose a system that can answer *why* questions by combining passages from different documents that answer the question. Even though the work is presented in the context of question answering, the output of the system is essentially a user-focused multi-document summary.

Personalised summaries are a specific type of user-focused summary where the summary is tuned to the needs of the user: Agnihotri, Kender, Dimitrova, and Zimmerman (2005) used personality tests to determine the users' interests to produce summaries of broadcast television content, whilst Díaz and Gervás (2007) produced summaries from newswire texts that are personalised based on a user profile which includes keywords, domain-specific factors and feedback received from the user. The background of a user is taken into consideration in *update summaries*, which are summaries that should focus only on information that is new to the user (Li, Liu, and Zhao, 2015). The interest in these types of summaries emerged largely as a result of the DUC and TAC conferences (see Section 4).

2.2.4 Indicative vs informative summaries

The literature also makes a distinction between *indicative summaries*, which give an indication of the contents of the source; *informative summaries*, which provide

much more information about the source, potentially replacing it; and *critical summaries*, which express the author's opinion on the information expressed in the source (Borko and Bernier, 1975; Mani, 2001b). The current methods developed in automatic texts summarisation are not really capable of producing "pure" indicative or informative summaries, instead producing something in between. The possible exceptions are perhaps the headline generation methods whose output is close to indicative summaries (Knight and Marcu, 2002; Rush, Chopra, and Weston, 2015). Critical summaries are currently beyond the capabilities of existing methods.

2.2.5 Language of summaries

Most of the early work in text summarisation focused on the processing of English texts. However, the need to process texts in other languages led researchers to produce not only *monolingual summaries*, where the input and the output are in the same language, but also *multilingual summaries*, where the input is in several languages and the output is in one of these languages, as well as *cross-lingual summaries*, where the language of the summary is different from the language of the input source(s) (Mani, 2001a). Multilingual and crosslingual summarisation methods usually involve some kind of translation engine (Orăsan and Chiorean, 2008; Wan, 2011).

3 A brief history of automatic text summarisation

This section presents a short history of the research carried out in automatic text summarisation. As can be seen below, the progress of the field and the approaches used were very much influenced by the paradigms of the time, moving back and forth between empirical and rationalist approaches.

3.1 The empiricism of the 1950s and 1960s

Research in automatic summarisation started in the early years of Artificial Intelligence (AI) which were described as days of "early enthusiasm, great expectations" by Russell and Norvig (2010). The first attempt to produce automatic summaries is credited to Luhn (1958), who noticed that statistical information derived from word frequencies can be used to determine the importance of sentences in a given text. The next significant publication was by Edmundson (1969), who pointed out that it is not enough to rely only on word frequencies for identifying the important sentences in texts. He proposed a number of other factors that should also be taken into consideration, such as presence of predefined cue words in a sentence that can boost or reduce the sentence's importance, whether the sentence contains words from the title, as well as the location of the sentence in the document or paragraph. The methods developed in this period produced extracts.

These two publications are very important for the field because the features employed by Luhn (1958) and Edmundson (1969) to identify the important sentences are still used in one way or another by current summarisation methods. In

addition, Edmundson (1969) is the first one to have assessed how different features influence the resulting summaries, a process which nowadays is used on regular basis in machine learning approaches. Despite the optimistic tone of both papers, the progress in the field was not as fast as it was hoped. The limitations imposed by the hardware available then were a key factor in this: Luhn (1958) mentions that all the texts had to be punched on cards before they were processed, whilst Edmundson (1969) could not process texts that had more than 4,000 words due to limited computer storage. In addition, Edmundson calculated that it cost approximately 1.5 cents per word to produce summaries. All these made it impossible to achieve the goal of producing summaries which "save a prospective reader time and effort in finding useful information in a given article or report" (Luhn, 1958).

3.2 The Rationalism of the 1970s and 1980s

The two publications mentioned above are essentially empirical approaches to producing summaries, whilst the 1970s and 1980s have marked a shift of research towards rationalist approaches. In the 1970s and 1980s, the main focus of research in artificial intelligence was on the development of methods which relied heavily on information about the problem to be resolved, and in some cases, tried to solve problems as a human would do it (Russell and Norvig, 2010). In many cases this information was domain-dependent, making the applicability of these methods somehow limited. These methods were also used in automatic summarisation with attempts being made to develop systems which "understand" the input texts and generate summaries on the basis of the understood information. For this reason, these methods are sometimes referred to as *understand and generate*. Typical examples of systems that use this approach are FRUMP (DeJong, 1982) and SUSY (Fum, Guida, and Tasso, 1985). FRUMP relied on manually created *sketchy scripts* to encode information extracted by rules from news stories as a way to "understand" them. The scripts were then used to generate summaries. SUSY tried to replicate the way humans summarise texts by trying to implement the theory proposed by Kintsch (1974). The approaches from this period were interesting, but their applicability was limited due to the fact that they were domain-dependent and relied on hand-coded information to encode rules and knowledge about the domain. Most of the methods produced *abstracts*, which means that the output contains units that are not present in the source in that form.

Despite the decline in the use of rationalist approaches at the beginning of the 1990s, researchers continued to develop methods inspired by this paradigm in order to produce abstracts. These approaches usually included a Natural Language Generation (NLG) module which outputs text from some form of internal representation (Reiter and Dale, 1997). The understanding of the source is usually achieved using robust information extraction methods like LaSIE (Gaizauskas and Humphreys, 1997) and InfoXtract (Srihari, Li, Cornell, and Niu, 2008) which fill in templates, rather than domain-specific extraction rules.

3.3 The Empiricism of 1990s to present

Twenty-five years ago, when the first issue of JNLE was published, the field of NLP was experiencing another paradigm shift, changing its focus from rationalist methods back to empirical/data-driven methods (Brill and Mooney, 1997). This shift was also reflected in the approaches used to produce summaries, and led to a renewed interest in automatic summarisation. As a result, researchers proposed new summarisation methods and developed better evaluation approaches (discussed in Section 4). The remainder of this section presents a brief description of the new methods proposed in the last 25 years and how their evolution can split into three overlapping periods. The section finishes with a brief description of summarisation-related fields which emerged since the first issue of JNLE.

3.3.1 The revival of empirical methods

The **first period** is characterised by methods which determine the importance of sentences either by using empirical observations about the properties of the input or by apply existing linguistic theories. Similarly to the rationalist approaches, this importance is determined using procedures and heuristics explicitly coded by researchers, in contrast to the methods presented in the Sections 3.3.2 and 3.3.3 where these are automatically derived from the data. For example, researchers employed methods from information retrieval which calculate links between texts and parts of texts (Salton, Singhal, Mitra, and Buckley, 1997) or relied on graph-based ranking models (Mihalcea and Tarau, 2004; Erkan and Radev, 2004) to identify the most appropriate sentences for a summary. Analysis of the anaphoric and coreferential links (Boguraev and Kennedy, 1999; Azzam, Humphrey, and Gaizauskas, 1999; Mitkov, Evans, Orăsan, Ha, and Pekar, 2007) or lexical repetition (Barzilay and Elhadad, 1999) in texts were also used to calculate a score for all the sentences in texts and extract only those with the highest score. Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), a theory which organises text in mainly non-overlapping spans linked by rhetorical relations, was successfully used to develop heuristics for extracting the most relevant sentences for a summary (Marcu, 1997, 2000; Alonso i Alemany and Fuentes Fort, 2003).

The main drawback of these methods is that they rely on researchers' intuitions about how to assess the importance of a sentence and use approximations to implement complex linguistic theories like RST. As a result, the summaries produced are not always good. These methods were preferred largely during the early stages of this re-emergence of the field and became less popular after the year 2000. However, successful applications of these approaches can be found even later on, for example in (Lloret and Palomar, 2013). In addition, these methods are still used to derive features for machine learning based summarisation approaches and methods such as TextRank (Mihalcea and Tarau, 2004) are still used as baselines.

The late 1990s saw the first workshops dedicated to the topic of automatic text summarisation: the *ACL'97/EACL'97 Workshop on Intelligent Scalable Text*

Summarization (ISTS)[4] held in Madrid, Spain in July 1997 and the *AAAI Spring Symposium on Intelligent Text Summarization*[5] organised in March 1998 in Palo Alto, California. As a result of the interest they received from the research community, workshops dedicated to automatic text summarisation are now organised on a regular basis at major conferences such as ACL, NAACL, EACL and RANLP.

TISTER Text Summarisation Evaluation (SUMMAC), the first evaluation conference dedicated to automatic summarisation, was completed in May 1998 (Mani, Firmin, House, Chrzanowski, Klein, Hirshman, Sundheim, and Obrst, 1998; Mani, Klein, House, Hirschman, Firmin, and Sundheim, 2002), and the first easily available corpora were developed in 2001 part of the Document Understanding Conferences (DUC) (see Section 4 for more details about these evaluation conferences). In addition to releasing annotated datasets, DUC have also contributed to the establishment of standard evaluation metrics. Once data became available to train machine learning algorithms, researchers experimented with the majority of machine learning methods in an attempt to produce original research and improve on the state of the art.

3.3.2 Machine learning based methods

The use of machine learning approaches to produce summaries is the main characteristic of the **second period**. The idea to train a classifier capable of identifying which sentences should be included in a summary was first used in (Kupiec et al., 1995), but it really started being employed on a regular basis only after annotated corpora became available. The corpora developed in the DUC, as well as the automatic evaluation metrics proposed in DUC, had the biggest impact on the community as they provided the means of comparing methods directly. Before the availability of these corpora, researchers had to develop their own annotated resources, which in many cases were specific to the research questions they were trying to answer (Teufel and Moens, 1997, 2002).

Over years researchers tried virtually every single machine learning algorithm available in an attempt to produce better summaries. The methods tested range from Bayesian classifiers (Kupiec et al., 1995; Teufel and Moens, 1997; Neto, Freitas, and Kaestner, 2002) and decision trees (Mani and Bloedorn, 1998; Neto et al., 2002; Knight and Marcu, 2002) to hidden Markov models (Conroy and O'leary, 2001) and integer linear programming (Luo, Liu, Liu, and Litman, 2018). Knight and Marcu (2002) adapt the noisy channel used in Statistical Machine Translation in order to develop a method for sentence compression which is seen as a first step towards producing summaries automatically. The summarisation process was also seen as an optimisation problem in (Naserasadi, Khosravi, and Sadeghi, 2019) where weights are learnt from the data.

The proliferation of machine learning and of ROUGE, the most used automatic

[4] `https://aclweb.org/anthology/volumes/intelligent-scalable-text-summarization/`
[5] `https://www.aaai.org/Library/Symposia/Spring/ss98-06.php`

evaluation metric in automatic summarisation (see Section 4 for more details), also meant that "literature is turning into a giant leaderboard, where publication depends on numbers and little else (such as insight and explanation)" (Church, 2017). This means that it is common that researchers report lots of numbers without trying to understand their meaning with respect to the automatic summaries evaluated. This made researchers seek more meaningful evaluation methods, which are discussed in Section 4.

3.3.3 Deep learning for text summarisation

Traditional machine learning approaches are still used, but they are slowly being replaced by methods based on deep learning. The introduction of neural approaches for text summarisation marked the beginning of the **third period** and happened around 2015. This mirrors the changes that have taken place in other fields in computational linguistics where the use of deep learning technologies have led to new and more accurate methods. Nowadays there are numerous papers where researchers try to apply the latest neural models for automatic text summarisation in an attempt to improve on the state of the art. In some cases the use of neural architectures is not entirely justified as the improvements are minimal, but when applied properly the new methods enable researchers to produce better summaries. In addition, the use of neural approaches revived the research on non-extractive summarisation and multimodal summarisation.

Most of the methods proposed in automatic summarisation were inspired by approaches first developed in neural machine translation (NMT). In contrast to NMT, and machine translation (MT) in general, the output of a summarisation system is much shorter than the source and does not necessarily depend on its length. This poses some challenges when adapting methods from NMT. In addition, the summarisation process leads to loss of information (ideally unimportant information), whereas MT should produce an accurate representation of the source without losing any information. This complicates the processing further, requiring that a way of determining the most important information in the source is integrated in the neural architecture.

Rush et al. (2015) present an attention-based summarisation method which is inspired by an attention-based encoder used in NMT (Bahdanau, Cho, and Bengio, 2014), but relies on a modified scoring function to ensure that it focuses on the most important information in the source. The method is used to generate headlines from newspaper articles, and is trained on pairs of headlines and first sentences from newspaper articles. For the same task, Nallapati, Zhou, dos Santos, Gulcehre, and Xiang (2016) develop an abstractive text summarisation method using Attentional Encoder-Decoder Recurrent Neural Networks which also builds on (Bahdanau et al., 2014). Filippova, Alfonseca, Colmenares, Kaiser, and Vinyals (2015) propose a Long Short Term Memory (LSTM) based method for sentence compression.

A sequence-to-sequence framework is used in (Cohan, Dernoncourt, Kim, Bui, Kim, Chang, and Goharian, 2018) to produce summaries of scientific documents. In contrast to the methods presented above, the output of this method is much

longer than those used for headline generation, making it a more difficult task. In addition, the encoder models the structure of the discourse of scientific documents in an attempt to produce better summaries.

Deep learning methods have been used also for extractive summarisation. Kobayashi, Noguchi, and Yatsuka (2015) and Yogatama, Liu, and Smith (2015) present two approaches that take advantage of the semantic information provided by word embeddings and propose unsupervised optimisation algorithms that can find the best set of sentences given the search space created by the semantic representation of the sentences in the document. In contrast to the methods mentioned above as well as the most methods proposed in this period, these two papers do not employ any neural network, but rely on word embeddings that are fundamental for any neural approach. Cheng and Lapata (2016) developed a framework for extractive summarisation which uses a neural network-based hierarchical document encoder and an attention-based content extractor. This framework is able to extract both sentences and words.

This section has presented only a small sample of the new methods which produce summaries with the help of neural networks. As it was the case with traditional ML-based methods, researchers try any new neural architecture that proves useful in other fields with the hope that it will lead to better summaries. For this reason, it is expected that this direction of research will still be active for years to come. One of the big challenges in using deep learning is the need for large datasets. Section 4.2 discusses how researchers have addressed this problem.

3.3.4 Summarisation-related fields

Advances in computational linguistics have led to the emergence of new fields which combine methods from automatic summarisation with other types of processing. This section describes a few such examples.

The growing interest in sentiment analysis and opinion mining (Liu, 2012) has led to the emergence of *sentiment-based summaries*, also referred to as *opinion summaries*, which attempt to capture the opinions in a large collection of documents such as online reviews (Carenini, Ng, and Pauls, 2006; Lerman, Blair-Goldensohn, and Mcdonald, 2009). These summaries are not necessarily coherent pieces of text. For example, they can be lists of features of the products reviewed with scores reflecting the number of positive/negative opinions about them (Balahur, Kabadjov, Steinberger, Steinberger, and Montoyo, 2012). At present, online shops (e.g. Amazon) and booking sites (e.g. hotels.com) summarise the main points of customer reviews in this format. Lerman, Lerman, McDonald, and McDonald (2009) discusses the possibility of producing *contrastive summaries* which highlight the sentiments of people about two different entities. As expected the success of these summarisation methods is influenced very much by the accuracy of the sentiment classifiers employed and how capable the methods are at extracting relevant features for products.

The combination of methods from automatic summarisation and citation analysis is commonly used to produce citation summaries and assess the impact of research.

Hernández-Alvarez and Gomez (2016) present a survey of the work carried out in the field of citation context analysis, whist Qazvinian and Radev (2008) propose a method that uses sentences containing citations occurring in scientific papers to a target paper to summarise this target paper. This summary is referred to as a *citation summary* and is an example of applying multi-document summarisation methods to a specific setting. Related to this are *survey summaries* which produce a summary about a topic or an entity starting from a biography (Zhou, Ticrea, and Hovy, 2004) or generate a related work section for a target article (Chen and Zhuge, 2016).

As mentioned above, text summarisation methods have been embedded in question answering systems in order to answer non-factoid questions (Verberne et al., 2010; Yang, Ai, Spina, Chen, Pang, Croft, Guo, and Scholer, 2016). The advances in computer vision means that multimodal summarisation is more feasible now, with systems that are able to caption images (Tanti, Gatt, and Camilleri, 2018) or are able to summarise complex sentences with images and other graphical representations (UzZaman, Bigham, and Allen, 2011). Methods from automatic summarisation also proved useful for text simplification (Margarido, Pardo, Antonio, Fuentes, Aires, Aluísio, and Fortes, 2008).

4 Evaluation

Evaluation in automatic summarisation is a very difficult task. The main challenge comes from the fact that there is no clear notion of what constitutes a good summary. For this reason it is challenging to define evaluation methods for automatic summarisation. In order to determine the quality of a summary, we have to consider several "fuzzy" factors such as whether it contains the important information from a document, omits unimportant information, does not contain redundant information, it presents information in a coherent and logical order, whether it is legible, and it is not misleading. All these notions are highly subjective and difficult to implement in programs. In addition, the context in which a summary is to be used should also be taken into account, because summaries which are good in one context could be inappropriate in another one. Another difficulty in summary evaluation comes from the fact that it is possible to produce more than one "correct summary" from a text, making it more difficult to define this notion.[6] Sparck-Jones (2001) points out that "a summary is a radical transformation of its source, implying far more possible output alternatives than in the relatively more limited MT situation."

The standard classification of the evaluation methods employed in automatic summarisation makes a distinction between intrinsic and extrinsic evaluations (Sparck-Jones and Galliers, 1996). If the results of a system are directly evaluated, the evaluation method is *intrinsic evaluation*, whereas *extrinsic evaluation* is

[6] Actually, the number of perfectly acceptable summaries that can be produced from a text is unlimited, due to the possibility of expressing a finite set of ideas in a virtually unlimited number of ways using lexical, syntactic and discourse variations.

performed when another system which uses the results of the first is evaluated, thereby taking into account the effect the results of the system under investigation have on another system. Hirschman and Mani (2003) point out that for intrinsic evaluation, measures such as quality and informativeness are recorded, whilst in extrinsic evaluation, post-edit measures, relevance assessment and reading comprehension tests are commonly used. A more detailed classification of types of evaluation methods used in automatic summarisation is discussed in (Tucker, 1999; Orăsan, 2006), but that classification is too specific for the purpose of this article. Although widely used, the appropriateness of intrinsic evaluation was questioned by Sparck-Jones (2001) because there is "nothing like a natural summary for a text". Instead she suggests considering the context and usage for summaries, which in most cases means performing extrinsic evaluation.

This section discusses the main issues involved in the evaluation of automatic text summarisation. Given how important evaluation conferences were for the development of the field, I start by presenting information about some of the most important evaluation conferences in Section 4.1. Extrinsic evaluations usually take a large amount of resources, both in terms of costs and time, in order to be carried out. For this reason, most of the extrinsic evaluations were carried out in the context of these evaluation conferences and will be mentioned as well. Intrinsic evaluations usually require annotated corpora. Considerations about how to build corpora for the field and brief descriptions of some of the existing corpora are presented in Section 4.2. Existing intrinsic evaluation methods used in automatic summarisation are briefly discussed in Section 4.3.

4.1 Evaluation conferences

Like in many other fields, research in automatic text summarisation was boosted by the organisation of evaluation conferences. The first such conference in automatic summarisation was TISTER Text Summarisation Evaluation (SUMMAC) completed in May 1998 (Mani et al., 1998, 2002). The purpose of SUMMAC was to organise the first large-scale, independent evaluation in automatic summarisation, in order to judge the participating systems "in terms of their usefulness in specific summarization tasks and to gain a better understanding of the issues involved in building and evaluating such systems" (Mani et al., 2002). SUMMAC allowed direct comparison between different systems for the first time, but did not attempt "to systematically classify the different technologies to study their effect on performance" (Mani et al., 2002). Even though SUMMAC included an intrinsic evaluation task where human judges were asked to evaluate the computer-generated summary in terms of informativeness, the main focus of the evaluation was on two extrinsic evaluation tasks: an *ad-hoc task*, in which indicative user-focused summaries were used to determine the relevance of the source document to a query, and a *categorisation task* in which indicative generic summaries were used to assign a document to a category.

One of the reasons SUMMAC focused on extrinsic evaluation is that it could model tasks of interest for the funding agencies and prove the usefulness of

summarisation in real scenarios. This echoes the point made by Sparck-Jones (2001) about the need to evaluate how a summary can be used. However, extrinsic evaluations are time-consuming and expensive to run. In addition, it is not possible to repeat them to assess incremental improvements of a system. For this reason, a significant amount of effort was dedicated to the development of the automatic evaluation methods presented in Section 4.3.

SUMMAC was followed by a series of Document Understanding Conferences (DUC) which focused on a number of increasingly difficult summarisation tasks, and employed largely intrinsic evaluation methods. A total of seven DUC evaluations were organised between 2001 and 2007. Over, Dang, and Harman (2007) present an overview of the DUC conferences, and general information about the conferences, the datasets and papers describing the participating systems can be found on DUC's website[7]. After the last Document Understanding Conference in 2007, the Text Analysis Conferences (TAC)[8] proposed a number of new summarisation tasks such as opinion summarisation and update summarisation (see Section 2), in addition to summarisation of documents in languages other than English. These days summarisation-related shared tasks are still organised, but they tend to be much more specialised which also means they attract fewer participants.

Overall it can be said that from all the evaluation conferences organised in the field so far, the Document Understanding Conferences have had the greatest impact as they created annotated corpora which are still being used by researchers, and they established the ROUGE method as the main intrinsic evaluation method used in the field.

4.2 Corpora for automatic summarisation

As in many other fields in computational linguistics, corpora play a very important role in automatic summarisation and are used both to train and evaluate summarisation methods. Corpora have also been employed in a limited number of cases to investigate features of summaries in order to design summarisation methods and to learn more about human summarisation. This latter research was carried out especially in the context of summarisation of scientific texts.

The first corpora for automatic summarisation had annotation which indicated the important sentences in texts (for example the CAST corpus (Hasler, Orăsan, and Mitkov, 2003)). This made them unsuitable for evaluating abstracts. They also had some serious limitations when it came to evaluating extracts because they expected a perfect match between the sentences extracted and those annotated, and they did not reward in any way when sentences with meaning similar to those annotated were extracted instead. Radev, Jing, and Budzikowska (2000) addressed this problem by using *utility judgement* which requires the assignment of a score from 0 to 10 for each sentence as to how relevant that sentence would be to a given

[7] `https://duc.nist.gov/`
[8] https://tac.nist.gov/tracks/index.html

topic. However, this process is not easy and therefore cannot be used on a large scale.

The Document Understanding Conferences (see Section 4.1) have developed alternative approaches to evaluate summaries which lessen the need for a distinction between abstracts and extracts, and also released gold standards that could be used by researchers. These gold standards usually consisted of texts accompanied by several human-produced "ideal" summaries. Even though developed more than 15 years ago, the DUC corpora are still being widely used in the field.

The deep learning methods presented in Section 3.3.3 require corpora that are bigger than those that can be annotated manually. For this reason, researchers have explored ways of producing such corpora automatically. Hermann, Kočiský, Grefenstette, Espeholt, Espeholt, Kay, Suleyman, and Blunsom (2015) developed two corpora by collecting articles from the CNN and Daily Mail websites. All the articles are accompanied by bullet points which summarise their context. The text in the bullet points is not directly extracted from the source documents which makes them appropriate for abstractive summarisation. Nowadays these corpora are regularly used to train deep learning based summarisation methods. For summarisation of scientific papers, Cohan et al. (2018) collected documents from scientific repositories such as arXiv.org and PubMed.com to build a corpus of scientific papers and their author-provided summary.

4.3 Intrinsic evaluation methods in text summarisation

.

Despite the reservations about intrinsic evaluation expressed in (Sparck-Jones, 2001), this type of evaluation is used extensively in automatic summarisation as it is usually simpler than extrinsic evaluation and can be applied by researchers independently of official evaluations. Intrinsic evaluation metrics require that human judges read and evaluate summaries according to some guidelines, or a (semi-)automatic method is used to compare a summary with a gold standard. The former can be quite difficult and expensive to run. In addition it makes evaluation of incremental improvements of the system difficult. Despite attempts to standardise it (Minel, Nugier, and Piat, 1997), this evaluation is rarely used these days.

The first automatic method employed to assess summaries compared a list of sentences selected by a summarisation method with a list of sentences considered the best sentences to extract from the source, and report the accuracy of the summarisation method using metrics such as precision and recall (Kupiec et al., 1995). This approach is not ideal because it is only appropriate for extracts and it does not allow any flexibility with respect to which sentences should be extracted.

The need to have a better way of calculating the informativeness of a summary has led to alternative ways of comparing the content of a summary with a gold standard. Donaway, Drummey, and Mather (2000) proposed using cosine similarity between an automatic summary and the gold standard to measure the information content. This approach works well, but it was superseded by ROUGE (Recall-Oriented Understudy for Gisting Evaluation), an evaluation metric

proposed in (Lin and Hovy, 2003; Lin, 2004) and inspired by BLEU, the standard metric used in machine translation evaluation (Papineni, Roukos, Ward, and Zhu, 2002). The assumption this method makes is that two texts have similar meanings if they share words and phrases. For this reason, ROUGE relies on the number of overlapping units such as n-grams, word sequences, and word pairs between an automatic summary and a human-produced summary, to assess the quality of the automatic summary. Lin and Hovy (2003) shows that ROUGE correlates well with human judgements. As a result, it was adopted as the main evaluation metric for DUC from 2004 and it has became the de facto evaluation metric in automatic summarisation.

Researchers have questioned whether ROUGE is really able to capture the informativeness of summaries. Over the years, other automatic metrics have been proposed like *Basic Elements* (Hovy, Lin, Zhou, and Fukumoto, 2006), AutoSummENG (Giannakopoulos and Karkaletsis, 2009) and GEMS (Generative Modelling for Evaluation of Summaries) (Katragadda, 2010) to name a few, but none managed to demonstrate enough advantages to replace ROUGE as the standard evaluation metric. Lloret and Palomar (2011) discuss in detail these alternative evaluation metrics, whilst Owczarzak, Conroy, Dang, and Nenkova (2012) provide an assessment of evaluation metrics used in multi-document summarisation evaluations.

The ROUGE metric does not necessarily capture the information content of a summary well enough. For this reason, (Nenkova and Passonneau, 2004; Passonneau, 2010) proposed the *pyramid method*, a semi-automatic evaluation measure which focuses on assessing the informativeness of summaries. This method assumes that summary content units (SCUs) can be manually identified in several human-produced summaries of the same text. The frequency of these SCUs is then used to assign them weights. An automatic summary, or any other text which was not used to identify the SCUs, is scored by summing the scores of the SCUs present in the them. Passonneau (2010) shows that it is possible to obtain high interannotator agreement when identifying SCUs and that the pyramid method is a good way to differentiate between summarisation systems. For this reason, the method was used in three DUC evaluations (2005 - 2007) and it is still used in some other evaluation conferences. Harnly, Nenkova, Passonneau, and Rambow (2005) propose an automatic method to identify SCUs, one of the main limitations of the pyramid method, but it is not widely used.

The fact that ROUGE can be calculated easily and the availability of corpora for summarisation has enabled researchers to develop and test new methods. A drawback of this is that it has became common that researchers only present the ROUGE scores without any attempt to check what the summaries look like.

5 Conclusions

Given that the amount of textual information available is already very large and that it will continue to increase, there will be a continuous need for summarisation. The existing systems are far from perfect, but they are already making a difference.

A number of language processing APIs already provide users with means to summarise their texts without the need to implement anything. In 2013, Yahoo acquired the news summarisation app *Summly* for an alleged 30 million US dollars. It is beyond the scope of this article to discuss whether the quality of the summaries produced by Summly was good enough or whether the app was really worth that much. However, this shows that there is a need for automatic summarisation in the commercial world and that the research that is being carried out is ready to be made into commercially viable applications.

The neural-based summarisation methods that have been proposed in the last five years have revived the field, especially by developing approaches that produce abstracts. I am confident that the coming years will see better neural-based summarisation methods that tackle further problems in the field.

In my opinion, at present the main obstacle faced by the field is the lack of adequate evaluation metrics. ROUGE is easy to apply and widely used, but even if we accept it as a good way of measuring the informativeness of a summary, which some researchers challenge, we still do not have widely accepted automatic methods for measuring the coherence and cohesion of summaries. As other NLP fields also progress, it may become more feasible to have fully automatic extrinsic evaluations of automatic summarisation, in this way providing further ways of assessing summarisation systems.

References

Agnihotri, L., J. Kender, N. Dimitrova, and J. Zimmerman (2005). User Study for Generating Personalized Summary Profiles. In *Procceings of the 2005 IEEE International Conference on Multimedia and Expo*, pp. 1094–1097. IEEE.

Alonso i Alemany, L. and M. Fuentes Fort (2003). Integrating cohesion and coherence for Automatic Summarization. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*.

Ando, R., B. Boguraev, R. Byrd, and M. Neff (2005, mar). Visualization-enabled multi-document summarization by Iterative Residual Rescaling. *Natural Language Engineering 11*(1), 67–86.

ANSI (1977, dec). American National standard for writing abstracts. *IEEE Transactions on Professional Communication PC-20*(4), 252–254.

Azzam, S., K. Humphrey, and R. Gaizauskas (1999). Using Coreference Chains for Text Summarisation. In A. Bagga, B. Baldwin, and S. Shelton (Eds.), *Proceedings of the Workshop on Coreference and its Applications*, Maryland, USA, pp. 77 – 84.

Bahdanau, D., K. Cho, and Y. Bengio (2014, sep). Neural Machine Translation by Jointly Learning to Align and Translate.

Balahur, A., M. Kabadjov, J. Steinberger, R. Steinberger, and A. Montoyo (2012). Challenges and solutions in the opinion summarization of user-generated content. *Journal of Intelligent Information Systems 39*, 375–398.

Barzilay, R. and M. Elhadad (1999). Using lexical chains for text summarization. In I. Mani and M. T. Maybury (Eds.), *Advances in automatic text summarization*, Chapter 10, pp. 111 – 122. The MIT Press.

Barzilay, R. and K. R. McKeown (2005, sep). Sentence Fusion for Multidocument News Summarization. *Computational Linguistics 31*(3), 297–328.

Boguraev, B. and C. Kennedy (1999). Salience-based content characterisation of text documents. In I. Mani and M. T. Maybury (Eds.), *Advances in automatic text summarization*, pp. 99 – 110. The MIT Press.

Borko, H. and C. L. Bernier (1975). *Abstracting concepts and methods*. Academic Press, London.

Brill, E. and R. J. Mooney (1997). An Overview of Empirical Natural Language Processing. *AI Magazine 18*(4), 13–24.

Carenini, G., R. Ng, and A. Pauls (2006). Multi-document summarization of evaluative text. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 305–312.

Chen, J. and H. Zhuge (2016, aug). Summarization of Related Work through Citations. In *2016 12th International Conference on Semantics, Knowledge and Grids (SKG)*, pp. 54–61. IEEE.

Cheng, J. and M. Lapata (2016). Neural Summarization by Extracting Sentences and Words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, pp. 484–494. Association for Computational Linguistics.

Church, K. W. (2017, may). Emerging trends: I did it, I did it, I did it, but. . . *Natural Language Engineering 23*(3), 473–480.

Cleveland, D. B. (1983). *Introduction to Indexing and Abstracting*. Libraries Unlimited, Inc.

Cohan, A., F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian (2018). A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana, USA, pp. 615–621. Association for Computational Linguistics.

Conroy, J. M. and D. P. O'leary (2001). Text summarization via Hidden Markov Models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, New Orleans, Louisiana, USA, pp. 406–407.

DeJong, G. (1982). An overview of the FRUMP system. In W. G. Lehnert and M. H. Ringle (Eds.), *Strategies for Natural Language Processing*, Book part (with own title) 5, pp. 149 – 177. Hillsdale, NJ: Lawrence Erlbaum.

Díaz, A. and P. Gervás (2007, nov). User-model based personalized summarization. *Information Processing & Management Management 43*(6), 1715–1734.

Donaway, R. L., K. W. Drummey, and L. A. Mather (2000). A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the NAACL-ANLP 2000 Workshop on Automatic summarization*, Morristown, NJ, USA, pp. 69–78. Association for Computational Linguistics.

Edmundson, H. P. (1969, apr). New Methods in Automatic Extracting. *Journal of the ACM 16*(2), 264–285.

Erkan, G. and D. R. Radev (2004). LexRank : Graph-based Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research 22*(1), 457–479.

Feiguina, O. and G. Lapalme (2007). Query-Based Summarization of Customer Reviews. In Z. Kobti and D. Wu (Eds.), *Advances in Artificial Intelligence. Canadian AI 2007*, pp. 452–463. Springer, Berlin, Heidelberg.

Filippova, K., E. Alfonseca, C. A. Colmenares, L. Kaiser, and O. Vinyals (2015). Sentence Compression by Deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 360–368. Association for Computational Linguistics.

Fum, D., G. Guida, and C. Tasso (1985). Evaluating importance: a step towards text summarisation. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, Los Angeles, California, pp. 840 – 844.

Gaizauskas, R. and K. Humphreys (1997). Using a semantic network for information extraction. *Natural Language Engineering 3*(2), 147–169.

Giannakopoulos, G. and V. Karkaletsis (2009). N-gram Graphs: Representing Documents and Document Sets in Summary System Evaluation. In *Proceedings of TAC 2009*.

Goldstein, J., M. Kantrowitz, V. Mittal, and J. Carbonell (1999). Summarizing text documents. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*, 10.1145/312624.312665, pp. 121–128. ACM Press.

Gupta, S. K. and S. K. Gupta (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications 121*, 49–65.

Harnly, A., A. Nenkova, R. Passonneau, and O. Rambow (2005). Automation of Summary Evaluation by the Pyramid Method. In *Proceedings of Recent Advances in Natural Language Processing*.

Hasler, L., C. Orăsan, and R. Mitkov (2003). Building better corpora for summarisation. In *Proceedings of Corpus Linguistics 2003*, Lancaster, UK, pp. 309 – 319.

Hermann, K. M., T. Kočiský, E. Grefenstette, L. Espeholt, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom (2015). Teaching Machines to Read and Comprehend NIPS 2015. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15)*, pp. 1693–1701.

Hernández-Alvarez, M. and J. M. Gomez (2016, may). Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering 22*(3), 327–349.

Hirschman, L. and I. Mani (2003). Evaluation. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*, Book part (with own title) 22. Oxford University Press.

Hovy, E. (2003). Text summarisation. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*, Book part (with own title) 32, pp. 583 – 598. Oxford University Press.

Hovy, E. and C.-Y. Lin (1999). Automated text summarization in SUMMARIST. In I. Mani and M. T. Maybury (Eds.), *Advances in automatic text summarization*, pp. 81 – 97. The MIT Press.

Hovy, E., C.-Y. Lin, L. Zhou, and J. Fukumoto (2006). Automated Summarization Evaluation with Basic Elements. In *Proceedings of the 5th international conference on language resources and evaluation (LREC)*.

Johnson, F. (1995, dec). Automatic abstracting research. *Library Review 44*(8), 28–36.

Katragadda, R. (2010). GEMS: Generative Modeling for Evaluation of Summaries. In *Computational Linguistics and Intelligent Text Processing. CICLing 2010*, pp. 724–735.

Kintsch, W. (1974). *The representation of meaning in memory*. The Experimental psychology series. Lawrence Erlbaum Associates Publishers.

Knight, K. and D. Marcu (2002, jul). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence 139*(1), 91–107.

Kobayashi, H., M. Noguchi, and T. Yatsuka (2015). Summarization Based on Embedding Distributions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 1984–1989. Association for Computational Linguistics.

Kupiec, J., J. Pedersen, and F. Chen (1995). A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '95*, Seattle, Washington, USA, pp. 68–73. ACM Press.

Lerman, K., S. Blair-Goldensohn, and R. Mcdonald (2009). Sentiment Summarization: Evaluating and Learning User Preferences. In *Proceedings of the 12th Conference ofthe European Chapter ofthe ACL (EACL 2009)*, Number April, Athens, Greece, pp. 514–522.

Lerman, K., K. Lerman, R. McDonald, and R. McDonald (2009). Contrastive summarization: an experiment with consumer reviews. In *Proceedings of NAACL HLT 2009: Short Papers*, Number June, Boulder, Colorado, pp. 113–116.

Li, C., Y. Liu, and L. Zhao (2015). Improving Update Summarization via Supervised ILP and Sentence Reranking. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA, pp. 1317–1322. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74 – 81.

Lin, C.-Y. and E. Hovy (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, Volume 1, Morristown, NJ, USA, pp. 71–78. Association for Computational Linguistics.

Liu, B. (2012, may). *Sentiment Analysis and Opinion Mining*, Volume 5.

Lloret, E. and M. Palomar (2011, apr). Text summarisation in progress: a literature review. *Artificial Intelligence Review 37*(1), 1–41.

Lloret, E. and M. Palomar (2013). COMPENDIUM: A text summarisation tool for generating summaries of multiple purposes, domains, and genres. *Natural Language Engineering 19*(2), 147–186.

Lloret, E., L. Plaza, and A. Aker (2018, mar). The challenging task of summary evaluation: an overview. *Language Resources and Evaluation 52*(1), 101–148.

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development 2*(2), 159–165.

Luo, W., F. Liu, Z. Liu, and D. Litman (2018, nov). A novel ILP framework for summarizing content with high lexical variety. *Natural Language Engineering 24*(6), 887–920.

Mani, I. (2001a). *Automatic Summarization*. Natural Language Processing. John Benjamins Publishing Company.

Mani, I. (2001b). Summarization evaluation: An overview. *Text*.

Mani, I. and E. Bloedorn (1998). Machine learning of generic and user-focused summarization. In *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence (AAAI '98/IAAI '98)*, pp. 820 –826. MIT Press.

Mani, I., T. Firmin, D. House, M. Chrzanowski, G. Klein, L. Hirshman, B. Sundheim, and L. Obrst (1998). The TIPSTER SUMMAC Text Summarisation Evaluation: Final Report. Technical report MTR 98W0000138, The MITRE Corporation.

Mani, I., G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim (2002, mar). SUMMAC: a text summarization evaluation. *Natural Language Engineering 8*(1), 43–68.

Mani, I. and M. T. Maybury (Eds.) (1999). *Advances in automatic text summarisation*. MIT Press.

Mann, W. C. and S. A. Thompson (1988). {Rhetorical Structure Theory: Toward a functional theory of text organisation}. *Text 3*(8), 234–281.

Marcu, D. (1997). From discourse structures to text summaries. In *Intelligent Scalable Text Summarization*, pp. 82 – 88.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarisation*. The MIT Press.

Margarido, P. R. A., T. A. S. Pardo, G. M. Antonio, V. B. Fuentes, R. Aires, S. M. Aluísio, and R. P. M. Fortes (2008). Automatic summarization for text simplification. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web - WebMedia '08*, New York, New York, USA, pp. 310. ACM Press.

Mihalcea, R. and P. Tarau (2004). TextRank: Bringing Order into Texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, pp. 404 – 411.

Minel, J. L., S. Nugier, and G. Piat (1997). How to appreciate the quality of automatic text summarization? In *Intelligent Scalable Text Summarization*, Madrid, Spain, pp. 25 – 30.

Mitkov, R., R. Evans, C. Orăsan, L. A. Ha, and V. Pekar (2007). Anaphora resolution: to what extent does it help NLP applications? In A. Branco (Ed.), *Lecture Notes In Artificial Intelligence*, pp. 179–190. Springer-Verlag.

Nallapati, R., B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang (2016). Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, Berlin, Germany, pp. 280–290. Association for Computational Linguistics.

Naserasadi, A., H. Khosravi, and F. Sadeghi (2019, jan). Extractive multi-document summarization based on textual entailment and sentence compression via knapsack problem. *Natural Language Engineering 25*(1), 121–146.

Nenkova, A. and K. McKeown (2012). A Survey of Text Summarization Techniques. In *Mining Text Data*, Chapter 3, pp. 43–76. Boston, MA: Springer US.

Nenkova, A. and R. Passonneau (2004). Evaluating content selection in summarization: The Pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 145 – 152.

Neto, J. L., A. A. Freitas, and C. A. A. Kaestner (2002). Automatic Text Summarization Using a Machine Learning Approach. In *Advances in Artificial Intelligence, 16th Brazilian Symposium on Artificial Intelligence*, Number November, pp. 205–215. Porto de Galinhas/Recife, Brazil.

Orăsan, C. (2006). *Comparative evaluation of modular automatic summarisation systems using CAST*. Phd thesis, University of Wolverhampton.

Orăsan, C. and O. A. Chiorean (2008). Evaluation of a Cross-lingual Romanian-English Multi-document Summariser. In *Proceedings of 6th Language Resources and Evaluation Conference (LREC2008)*, Marrakech, Morocco, pp. 2114 –2119.

Over, P., H. Dang, and D. Harman (2007, nov). DUC in context. *Information Processing & Management 43*(6), 1506–1520.

Owczarzak, K., J. M. Conroy, H. T. Dang, and A. Nenkova (2012, jun). An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, Montreal, Canada, pp. 1–9.

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2002, jul). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics Annual Meeting (ACL)*, Philadelphia, Pennsylvania, pp. 311 – 318.

Passonneau, R. J. (2010, apr). Formal and functional assessment of the pyramid method for summary content evaluation. *Natural Language Engineering 16*(2), 107–131.

Qazvinian, V. and D. R. Radev (2008). Scientific paper summarization using citation summary networks. In *Proceedings ofthe 22nd International Conference on Computational Linguistics (Coling 2008)*, Number August, Manchester, UK, pp. 689–696.

Radev, D. R., H. Jing, and M. Budzikowska (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In *Proceedings of the NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Reiter, E. and R. Dale (1997, mar). Building applied natural language generation systems. *Natural Language Engineering 3*(1), 57 – 87.

Rush, A. M., S. Chopra, and J. Weston (2015). A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 379–389. Association for Computational Linguistics.

Russell, S. and P. Norvig (2010). *Artificial Intelligence: A Modern Approach* (Third ed.). Pearson.

Saggion, H. (2008). Automatic summarization: an overview. *Revue française de linguistique appliquée 13*(1), 63–81.

Salton, G., A. Singhal, M. Mitra, and C. Buckley (1997). Automatic text structuring and summarization. *Information Processing and Management 33*(2), 193–207.

Spärck-Jones, K. (1999). Automatic summarizing: factors and directions. In I. Mani and M. T. Maybury (Eds.), *Advances in automatic text summarization*, Chapter 1, pp. 1–12. The MIT Press.

Sparck-Jones, K. (2001). Automatic language and information processing: rethinking evaluation. *Natural Language Engineering 7*(01), 29–46.

Sparck-Jones, K. and J. R. Galliers (1996). *Evaluating natural language processing systems: an analysis and review.* Number 1083 in Lecture Notes in Artificial Intelligence. Springer.

Srihari, R. K., W. Li, T. Cornell, and C. Niu (2008, jan). InfoXtract: A customizable intermediate level information extraction engine. *Natural Language Engineering 14*(01).

Tanti, M., A. Gatt, and K. P. Camilleri (2018, may). Where to put the image in an image caption generator. *Natural Language Engineering 24*(3), 467–489.

Teufel, S. and M. Moens (1997). Sentence extraction as a classification task. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, pp. 58 – 65.

Teufel, S. and M. Moens (2002, dec). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics 28*(4), 409–445.

Tigelaar, A. S., R. Op Den Akker, and D. Hiemstra (2010, apr). Automatic summarisation of discussion fora. *Natural Language Engineering 16*(2), 161–192.

Tucker, R. (1999). *Automatic summarising and the CLASP system.* Phd thesis, University of Cambridge, UK.

UzZaman, N., J. P. Bigham, and J. F. Allen (2011). Multimodal summarization of complex sentences. In *Proceedings of the 15th international conference on Intelligent user interfaces - IUI '11*, New York, New York, USA, pp. 43. ACM Press.

Verberne, S., L. Boves, N. Oostdijk, and P.-A. Coppen (2010, jun). What Is Not in the Bag of Words for Why -QA? *Computational Linguistics 36*(2), 229–245.

Verberne, S., E. Krahmer, S. Wubben, and A. van den Bosch (2019). Query-based summarization of discussion threads. *Natural Language Engineering* (May), 1–27.

Wan, X. (2011). Using bilingual information for cross-language document summarization. In *Proceedings ofthe 49th Annual Meeting ofthe Association for Computational Linguistics*, pp. 1546–1555.

Yang, L., Q. Ai, D. Spina, R.-C. Chen, L. Pang, W. B. Croft, J. Guo, and F. Scholer (2016). Beyond Factoid QA: Effective Methods for Non-factoid Answer Sentence Retrieval. In *Advances in Information Retrieval. ECIR 2016*, pp. 115–128.

Yao, J.-g., X. Wan, and J. Xiao (2017, nov). Recent advances in document summarization. *Knowledge and Information Systems 53*(2), 297–336.

Yogatama, D., F. Liu, and N. A. Smith (2015). Extractive Summarization by Maximizing Semantic Volume. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 1961–1966. Association for Computational Linguistics.

Zhou, L., M. Ticrea, and E. Hovy (2004). Multi-document Biography Summarization. *Proceedings of the Empirical Methods in Natural Langauge Processing (EMNLP 2004)*, 434–441.